CarbonData Home

Welcome to the Apache CarbonData wiki. If you are interested in contributing to CarbonData, visit the contributing to CarbonData page to learn more.

Release plan(Around 3 months for one release)

Date	Version number
Aug 2016	Apache CarbonData 0.1.0-incubating
Sep 2016	Apache CarbonData 0.1.1-incubating
Nov 2016	Apache CarbonData 0.2.0-incubating
Jan 2017	Apache CarbonData 1.0.0-incubating
May 2017	Apache CarbonData 1.1.0
Aug-Sep 2017	Apache CarbonData 1.2.0
Jan-Feb 2018	Apache CarbonData 1.3.0
Mar 2018	Apache CarbonData 1.3.1
May 2018	Apache CarbonData 1.4.0
Aug 2018	Apache CarbonData 1.4.1
Oct 2018	Apache CarbonData 1.5.0
Dec 2018	Apache CarbonData 1.5.1
Jan 2019	Apache CarbonData 1.5.2
Mar 2019	Apache CarbonData 1.5.3
May 2019	Apache CarbonData 1.5.4
Aug 2019	Apache CarbonData 1.6.0
Oct 2019	Apache CarbonData 1.6.1
May 2020	Apache CarbonData 2.0.0
Jun 2020	Apache CarbonData 2.0.1
Nov 2020	Apache CarbonData 2.1.0
Mar 2021	Apache CarbonData 2.1.1
Aug 2021	Apache CarbonData 2.2.0
Jan 2022	Apache CarbonData 2.3.0

Road map plan:

1.0.x:

- Support 2.1 integration in CarbonData
- Remove kettle, support new data load solution
- Support data update and delete SQL in Spark 1.6

1.1.x:

- Add page in blocklet for improving scan cases' performance.
- Support V3 format for improving TPC-H performance.
- Support vector features by default
- Support data update and delete SQL in Spark 2.1

1.2.x

- Support to specify sort column for MDK(Multi-Dimension Key index)

- Support partition
- Support Presto integration
- Support Hive integration
- Data loading optimization using ColumnPage in write step and make it off heap

1.3.x:

- Support streaming ingestion data to CarbonData
- Provide index framework for supporting user to add more index.
- Support local dictionary
- Ecosystem integration(eg. latest Apache Spark version 2.x)

1.4.x:

- Support create carbondata on cloud storage(AWS S3, Huawei OBS)
- Provide index framework for supporting user to add more index, like : text index using lucene
- Ecosystem integration

1.5.x:

- Support MV(Materialized View), Bloom Filter (in production features)
- Support CarbonData engine for improving concurrent visit and point query.
- Ecosystem integration
- Support alter add column in carbon file format
- Supports multiple character separators in csv file during data loading
- Compaction support for segments created with range_sort and global_sort
- Support DDLs to operate on Driver Cache (Get cache size, clear cache)
- Support building datamaps and data load in parallel to reduce the overall time taken
- Summary of loaded and bad records data after data loading

1.6.x:

- Support storing of carbon Min-Max indexes in external system
- MV DataMap Enhancements and Stabilisation
- Query performance enhancements
- Deeper Presto integration and stabilisation
- UDF and UDAF support in Pre-aggregate tables
- Support Read from hive

2.0.x:

- Support Write into hive
- Load performance improvements
- TPCDS [Query, load] performance improvements
- Carbon Advisor for auto suggestion of ideal table schema including MV, index, sort col, range col, compression ...
- Delete and update support in CarbonData SDK
- Support C engine reader for CarbonData SDK
- ES based datamap management
- Support Spark DataSource API V2
- Support CarbonData metadata management using DB or other external OLTP system
- Support MV on Streaming tables, partition tables, Time Series
- Support MV creation from another MV

2.1.x:

- Presto read support for complex columns
- Make GeoID visible to the user
- Support Carbondata SDK to load data from parquet, ORC, CSV, Avro and JSON.
- Implement delete and update feature in carbondata SDK.
- Support array<string> with SI
- Support IndexServer with Presto Engine
- Implementing a new Reindex command to repair the missing SI Segments
- Support Change Column Comment
- Support Local dictionary for presto complex datatypes
- Block Pruning for geospatial polygon expression
- Improve concurrent query performance
- Support global sort for Secondary index table
- Filter reordering
- Geospatial index algorithm improvement and UDFs enhancement
- CarbonData Trash support
- Support Writing Flink Stage data into Hdfs file system
- Support MERGE INTO SQL Command
- Support Complex DataType when Save DataFrame
- Adding global sort support for SI segments data files merge operation.

2.2.x:

- Support Add, Drop and rename column support for the complex column
- Spark-3.1 support
- Secondary Index Support for Presto
- CDC Performance improvement
- Local sort Partition Load and Compaction improvement
- Geo Spatial Query enhancements
- Improve table status and metadata writing

2.3.x:

- Support spatial index creation using data frame
- Introduce Streamer tool for Carbondata
- Upgrade prestosql to 333 version
- Multi-level complex schema support
- Support for Dynamic Partition Pruning

Pages Link

Committers

Releases

CarbonData Performance Reports

Apache CarbonData Performance Benchmark(0.1.0)

Events(Summit and Meetup materials)

Use cases and shared articles